

## Additional file 4. Equivalent numbers of independent loci

### Estimation of equivalent number of independent markers

Let  $\mathbf{x}_c^i$  and  $\mathbf{X}_r^i$  be the genotype matrices of the independent loci. We search the equivalent number of independent loci which gives the precision  $E[r_{q_c, \hat{q}_c}^2] \sim \frac{E_X[v(\hat{q}_c|\mathbf{X})]}{E_X[v(q_c|\mathbf{X})]}$  obtained with the total number of non-independent markers. It will be supposed that  $E_X[v(q_c|\mathbf{X})] = E[v(q_c)] = \sigma_q^2$  is the same considering all markers or only the set of independent loci. This may be a critical hypothesis if the correlation between the molecular score ( $q_c$ ) and the genetic value ( $g_c$ ) is not high. Under this hypothesis,  $\sigma_q^2 = \tau\sigma_\beta^2 = \tau_i\sigma_{i\beta}^2$  and the ratios  $\lambda_\beta = \frac{\tau}{\gamma}$  and  $\lambda_\beta^i = \frac{\tau_i}{\gamma}$ .

The equivalence will be observed if  $E_{\mathbf{x}_c, \mathbf{X}_r}[v(\hat{q}_c|\mathbf{x}_c, \mathbf{X}_r)] = E_{\mathbf{x}_c^i, \mathbf{X}_r^i}[v(\hat{q}_c|\mathbf{x}_c^i, \mathbf{X}_r^i)]$ . Using  $v(\hat{q}_c|\mathbf{x}_c, \mathbf{X}_r) = \sigma_\beta^2 \mathbf{x}_c \mathbf{x}_c' - \sigma_e^2 \mathbf{x}_c (\mathbf{X}_r' \mathbf{X}_r + \lambda_\beta \mathbf{I})^{-1} \mathbf{x}_c'$  and assuming that  $\mathbf{X}_r' \mathbf{X}_r \sim E[\mathbf{X}_r' \mathbf{X}_r]$  we have

$$E_{\mathbf{x}_c, \mathbf{X}_r}[v(\hat{q}_c|\mathbf{x}_c, \mathbf{X}_r)] = \sigma_\beta^2 E[\mathbf{x}_c' \mathbf{x}_c] - \sigma_e^2 \text{tr} \left[ (E[\mathbf{X}_r' \mathbf{X}_r] + \lambda_\beta \mathbf{I})^{-1} E[\mathbf{x}_c' \mathbf{x}_c] \right]$$

If  $E[\mathbf{x}_c' \mathbf{x}_c] = E[\mathbf{X}_r' \mathbf{X}_r]$ , the trace becomes  $\text{tr}[\mathbf{I}] - \lambda_\beta \text{tr} \left[ (E[\mathbf{X}_r' \mathbf{X}_r] + \lambda_\beta \mathbf{I})^{-1} \right]$

These relations hold for the independent loci case and the equivalence supposes that

$$-n_M \sigma_e^2 + \lambda_\beta \sigma_e^2 \text{tr} \left[ (E[\mathbf{X}_r' \mathbf{X}_r] + \lambda_\beta \mathbf{I})^{-1} \right] = -n_{M_i} \sigma_e^2 + \lambda_\beta^i \sigma_e^2 \text{tr} \left[ (E[\mathbf{X}_r^i' \mathbf{X}_r^i] + \lambda_\beta^i \mathbf{I})^{-1} \right] \quad [SM4.1]$$

The expectations matrices depend on marker alleles frequencies and linkage disequilibrium:

$$E[\mathbf{X}_r^i' \mathbf{X}_r^i] = n_R \begin{pmatrix} s_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_{n_{M_i}}^2 \end{pmatrix} \text{ and } E[\mathbf{X}_r' \mathbf{X}_r] = n_R \begin{pmatrix} \sigma_1^2 & \Delta_{12} & \cdots & \Delta_{1n_M} \\ \Delta_{21} & \sigma_2^2 & \cdots & \Delta_{2n_M} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{n_{M1}} & \Delta_{n_{M2}} & \cdots & \sigma_{n_M}^2 \end{pmatrix}$$

If we assume all genotypic values variances equal in the independent loci case, *i.e.*  $s_m^2 = s^2$ , it comes that  $\tau_i = n_{M_i} s^2$  and  $\text{tr} \left[ (E[\mathbf{X}_r^i' \mathbf{X}_r^i] + \lambda_\beta^i \mathbf{I})^{-1} \right] = \frac{n_{M_i}}{n_R s^2 + \lambda_\beta^i} = \frac{1}{s^2} \frac{\gamma n_{M_i}}{\gamma n_R + n_{M_i}}$ . The right hand side of [SM4.1] is

$$-n_{M_i} \sigma_e^2 + \frac{n_{M_i} s^2}{\gamma} \sigma_e^2 \frac{1}{s^2} \frac{\gamma n_{M_i}}{\gamma n_R + n_{M_i}} = -n_{M_i} \sigma_e^2 \left( 1 - \frac{n_{M_i}}{\gamma n_R + n_{M_i}} \right) = -n_{M_i} \sigma_e^2 \frac{\gamma n_R}{\gamma n_R + n_{M_i}}$$

Let  $T = \text{tr} \left[ (E[\mathbf{X}_r' \mathbf{X}_r] + \lambda_\beta \mathbf{I})^{-1} \right]$ , the equivalence gives  $-n_M + \lambda_\beta T = -n_{M_i} \frac{\gamma n_R}{\gamma n_R + n_{M_i}}$

Thus  $(-n_M + \lambda_\beta T)(\gamma n_R) = n_{M_i}(-\gamma n_R + n_M - \lambda_\beta T)$ , and using  $\tau = \sum_m \sigma_m^2 = n_M \overline{\sigma_m^2}$ , we get

$$n_{M_i} = \frac{(\overline{\sigma_m^2} T - \gamma)(\gamma n_R)}{(\gamma - \overline{\sigma_m^2} T - \gamma^2 n_R / n_M)} \quad [SM4.2]$$

In models predicting genetic gain expected from a genomic selection scheme applied to a real population, one of the parameters needed is the equivalent number of loci which can be estimated from formula [SM4.2] given marker alleles frequencies and between markers linkage disequilibrium observed in the population.

A very simple approximation for  $n_{M_e}$  is obtained assuming equal variances  $\sigma_m^2 = s^2$ , and using relation between expected linkage disequilibrium and effective population size  $N_e$  as derived by [32]  $E[2\Delta_{ml}] = \sigma_m \sigma_l / \sqrt{1 + 4N_e d_{lm}}$  with  $d_{lm}$  the distance between ordered locus  $l$  and  $m$ , such that  $d_{lm} = |l - m|L/n_M$ , with  $L$  the genome length in Morgans. With those hypotheses,  $T = \text{tr} \left[ (n_R s^2 \mathbf{R} + \lambda_\beta \mathbf{I})^{-1} \right]$  with  $\{\mathbf{R}\}_{ml} = r_{ml} = \sqrt{n_M / (n_M + 4N_e |l - m|L)}$ .

As  $\lambda_\beta = \frac{n_M s^2}{\gamma}$  we have  $T = \frac{\gamma}{s^2} \text{tr} [(\gamma n_R \mathbf{R} + n_M \mathbf{I})^{-1}] = \frac{\gamma}{s^2} U$

In this simplified situation, the equivalent number of loci is

$$n_{M_i} = n_M \frac{n_R \gamma (1 - U)}{n_R \gamma - n_M (1 - U)} \quad [SM4.3]$$

### Relation between trace T and the reference population size $n_R$ .

The objective is to understand how the number of independent markers  $n_{M_i}$  varies with  $n_R$ .

From [SM4.2], the derivative

$$\frac{dn_{M_i}}{dT} \propto \overline{\sigma_m^2} (\gamma n_R) \left( \gamma - \overline{\sigma_m^2} T - \frac{\gamma^2 n_R}{n_M} \right) - \overline{\sigma_m^2} (\overline{\sigma_m^2} T - \gamma) (\gamma n_R) = -\overline{\sigma_m^2} \gamma^3 \frac{n_R^2}{n_M} < 0$$

Thus, the  $n_{M_i}$  number decreases when the trace T increases.

In the simplified situation,  $T = \text{tr} \left[ (n_R s^2 \mathbf{R} + \lambda_\beta \mathbf{I})^{-1} \right]$  with  $r_{ml} = \sqrt{n_M / (n_M + 4N_e |l - m|L)}$ .

We consider the simplest situation of 2 markers. In this case,  $T = \frac{2(n_R s^2 + \lambda_\beta)}{(n_R s^2 + \lambda_\beta)^2 - (n_R s^2 r_{12})^2}$  showing that

(i) the trace under linkage disequilibrium is larger than the trace for independent markers ( $T_{LD} > T_{LE}$ ), (ii) the derivative of  $T_{LE}$  is negative ( $\frac{dT_{LE}}{dn_R} \propto -2s^2$ ) meaning that with larger  $n_R$ ,  $T_{LE}$  is lower

and  $n_{M_i}$  higher and (iii) the derivative of ratio  $T_{LE}/T_{LD} = 1 - \frac{(n_R s^2 r_{12})^2}{(n_R s^2 + \lambda_\beta)^2}$  is negative ( $\frac{dT_{LE}/T_{LD}}{dn_R} =$

$-2 \frac{n_R s^4 r_{12} \lambda_\beta}{(n_R s^2 + \lambda_\beta)^3}$ ) showing that the difference between those traces decreases when  $n_R$  becomes larger.

Globally, these information indicate that with larger reference population sizes, larger number of independent loci (figures closer to the number of markers in LD) are needed to obtain the same reliability.